Дисперсия – мера разброса значений, т.е.отклонение от выборки д-х.

**Анализ главных компонент** представляет собой метод, который осущ.вращение д-х с тем, чтобы преобразованные признаки не коррелировали между собой. Часто это вращение сопровождается выбором подмн-ва новых признаков в завис-ти от их важности с т.зр. интерпретации д-х.

[1] демонстрирует результат выполнения PCA к синтетическому 2мерному массиву д-х.

**1ый рисунок** (вверху слева) показывает исх.т.д-х, кот.выделены цветом для лучшей дискриминации. Алгоритм начинает работу с того, что сначала находит направление максимальной дисперсии, помеченное как «компонента 1». Речь идет о направлении д-х, кот.содержит большую часть инфы, т.е.находим признаки которые коррелируют друг с другом сильнее всего. Направления, которые найдены с помощью этого алгоритма называются **главными компонентами**, т.к.они являются основными направлениями дисперсии д-х. в целом максимально возможное кол-во главных компонент равно кол-ву исх.признаков.

1 – исз

**2ой график** (вверху справа) показывает те же самые д-е, но теперь повернуты таким образом, что первая компонента совпадает с осью х, а вторая главная компонента с осью у. перед вращением из каждого значения данных вычитается среднее, т.о.преобразованные данные центрированы около нуля. В новом представлении оси становятся некоррелированными. Т.е. все элементы корреляционной матрицы д-х, кроме диагональных будут равны 0.

Также можно использовать анализ главных компонент(РСА) для уменьшения размерности, сохранив лишь несколько компонент.

**3ий график** оставим лишь первую главную компоненту. Это уменьшит размерность д-х: из 2мерного массива д-х получаем 1мерный массив д-х. Однако, вместо того, чтобы оставить лишь 1 из исходных признаков, мы находим наиболее интересное направление (выходящее из верхнего левого угла в нижний правый на первом графике) и оставляем это направление, т.е.первую главную компоненту.

Можем отменить вращение и добавить обратно среднее значение к значениям д-х. в итоге получим данные, показанные на **4ом графике**. Эти точки располагаются в пр-ве исх. Признаков, но мы оставили лишь инфу, кот.содержится в первой компоненте. Это преобразование используется для того, чтобы удалить эффект шума из данных или показать какая часть инфы сохраняется при использовании главных компонент.

ПРИМЕНЕНИЕ РСА К НАБОРУ ДАННЫХ

Можно визуализировать применение РСА вычислив гистограммы распределения значений признаков для 2 классов, доброкачественных и злокачественных опухолей [2]

В данном случае мы строим для каждого признака гистограмму, подсчитывая частоту встречаемости точек данных в пределах границ интервалов. Интервал называют еще **бином.** Каждый график содержит 2 наложенные друг на друга гистограммы, первая – для всех точек класса доброкачественная опухоль, а вторая – для всех точек класса злокачественная опухоль. Это дает нам некоторое представление о распределении каждого признака по 2 классам и позволяет строить предположения о том, какие признаки лучше всего дискриминируют злокачественные и доброкачественные опухоли.

Например признак «smoothness error» довольно малоинформативен, потому что 2 гистограммы, построенные для данного признака, большей частью накладываются друг для друга. В то время как признак «worst concave points» кажется весьма ринформативным, поскольку гистограммы практически не накрывают друг друга.

Перед тем, как применить PCA можно отмасштабировать данные таким образом, чтобы каждый признак имел единичную дисперсию, воспользовавшись StandartScaler [3]

По умолчанию РСА лишь поворачивает д-е, но сохраняет все главные компоненты. Чтобы уменьшить размерность данных, нам нужно указать, сколько компонент мы хотим сохранить при создании объекта РСА [4]. Мы создаем экземпляр объекта РСА, находим главные компоненты, вызвав метод fit, а затем применяем вращение и снижение размерности, вызвав метод transform.

Теперь мы можем построить график первых 2 компонент [5]

PCA является методом машинного обучения без учителя и не использует какой-либо инфы о классах при поиске поворота. Он просто анализирует корреляционные связи в данных. Для точечного графика, показанного здесь, мы построили график, где по оси x отложена первая главная компонента, по оси y – вторая главная компонента, а затем воспользовались информацией о классах, чтобы выделить точки разным цветом. Вы можете увидеть, что в рассматриваемом двумерном пространстве эти два класса разделены достаточно хорошо.

Недостаток PCA заключается в том, что эти две оси графика часто бывает сложно интерпретировать. Главные компоненты соответствуют направлениям данных, поэтому они представляют собой комбинации исходных признаков.Сами главные компоненты могут быть сохранены в атрибуте components\_ объекта PCA в ходе подгонки [6].

Каждая строка в атрибуте components\_ соответствует одной главной компоненте и они отсортированы по важности (первой приводится первая главная компонента и т.д.). Столбцы соответствуют атрибуту исходных признаков для объекта PCA в этом примере, «mean radius», «mean texture» и т.д. Давайте посмотрим на содержимое атрибута components\_[7]

Кроме того, с помощью тепловой карты [8] мы можем визуализировать коэфф, чтобы упростить их интерпретацию. Можно увидеть, что в первой компоненте коэф.всех признаков имеют одинаковый знак. Это означает, что существует общая корреляция между всеми признаками. Высоким значениям одного признака будут соответствовать высокие значения остальных признаков. Во второй компоненте коэфициенты признаков имеют различные знаки. Обе компоненты включают все 30 признаков. Смешивание признаков – как раз то, что усложняет интерпретацию осей.

МЕТОД СОБСТВЕННЫХ ЛИЦ ДЛЯ ВЫДЕЛЕНИЯ ХАРАКТЕРИСТИК

Еще одно применение PCA, о котором мы уже упоминали ранее, – это выделение признаков. Идея, лежащая в основе выделения признаков, заключается в поиске нового представления данных, которое в отличие от исходного лучше подходит для анализа.

Мы приведем очень простой пример того, как можно применить выделение признаков к изображениям с помощью PCA. Для этого мы воспользуемся набором данных Labeled Faces in the Wild. Этот набор данных содержит изображения лиц знаменитостей, загруженных из Интернета, и включает в себя лица политиков, певцов, актеров и спортсменов с начала 2000-х годов. Мы уменьшим их для более быстрой обработки. [9].

Получаем 3023 изображения 100\*75 пикселей, принадлежащие 62 различным людям [10].

Однако данные немного асимметричны. Как вы можете здесь увидеть, он содержит большое количество изображений Джорджа Буша и Колина Пауэлла [11].

Чтобы данные стали менее асимметричными, мы будем рассматривать не более 50 изображений каждого человека (в противном случае выделение признаков будет перегружено большим количеством изображений Джорджа Буша) [12]

Общая задача распознавания лиц заключается в том, чтобы спросить, не принадлежит ли незнакомое фото уже известному человеку из базы данных. Построим классификатор одного ближайшего соседа, который ищет лицо, наиболее схожее с классифицируемым. Этот классификатор в принципе может работать только с одним обучающим примером в классе. Давайте посмотрим, насколько хорошо здесь сработает KNeighborsClassifier [13]. Мы получаем правильность 26.6%, на самом деле это неплохо для классификационной задачи с 62 классами, но и не так велико. Мы правильно распознаем лишь каждое четвертое изображение человека.

И вот именно здесь применяется PCA. Мы рассчитываем на то, что использование расстояний вдоль главных компонент может улучшить правильность. Здесь мы воспользуемся опцией PCA выбеливание ( whitening ), которая преобразует компоненты к одному и тому же масштабу. Операция выбеливания аналогична применению StandardScaler после преобразования. Повторно используя данные выбеливание не только поворачивает данные, но и масштабирует их таким образом, чтобы центральный график представлял собой окружность вместо эллипса [14].

Мы подгоняем объект PCA на обучающих данных и извлекаем первые 100 главных компонент. Затем мы преобразуем обучающие и тестовые данные [15].

Новые данные содержат 100 новых признаков, первые 100 главных компонент. Теперь мы можем использовать новое представление, чтобы классифицировать наши изображения, используя классификатор одного ближайшего соседа [16]. Наша правильность улучшилась это подтверждает наше предположение о том, что главные компоненты могут дать лучшее представление данных.

Давайте посмотрим на первые несколько главных компонент [17]

Мы можем выполнить аналогичное преобразование для лиц, сократив данные за счет использования лишь некоторых главных компонент и вернувшись затем в исходное пространство. Это возвращение в пространство исходных признаков можно выполнить с помощью метода inverse\_transform. Здесь мы визуализируем результаты реконструкции некоторых лиц, используя 10, 50, 100, 500 и 2000 компонент [18]

Кроме того, мы можем применить PCA для визуализации всех лиц набора на диаграмме рассеяния, воспользовавшись первыми двумя главными компонентами (рис. 3.12). Для этого мы выделим классы, соответствующие лицам, с помощью определенного цвета и формы (аналогично тому, что делали для набора данных cancer) [19]

Из рис. видно, когда мы используем лишь первые две главные компоненты, все данные представляют собой просто одно большое скопление данных без видимого разделения классов. Данный факт неудивителен, учитывая, что даже при использовании 10 компонент PCA фиксирует самые общие характеристики лиц.